Real-Time Estimation of Propagation Delays for Temporal Alignment of Audio Signals in Augmented Reality Applications

Marcel Nophut[1], Robert Hupke[1], Stephan Preihs[1], Jürgen Peissig[1]
[1] *Leibniz Universität Hannover, Institut für Kommunikationstechnik, 30167 Hannover, Germany*
Email: nophut@ikt.uni-hannover.de

Abstract

In augmented reality audio applications a superposition of environmental sounds and supplementary audio content is used to create auditory enhancements for the listener in a broad range of use cases. In some use cases environmental sounds and supplementary content may be highly correlated, for example in audience services at live events, where a live playback through PA speakers is enhanced by augmented reality audio content, e.g. to create an individualized live mix. Without temporal alignment of those signals, a superposition can cause comb filtering effects or confusing echoes.

This contribution proposes an efficient method that is able to robustly detect a temporal offset of correlated audio signals. It is based on a recursive cross-correlation estimation and a peak detection algorithm. The method focuses on indoor music and speech events with their typically occurring problems like room reflections, tonal components and a large number of correlation lags. The temporal offset obtained is used to delay the supplementary audio content in order to achieve a temporal alignment of the signals.

Introduction

In the course of the project PMSE-xG [1], a novel audience service, the “Assistive Live Listening Service”, was developed [2]. This service provides individualized audio content to listeners at a concert or voice-based live event. It can be used to achieve enhancements of the listening experience, e.g. to overcome poor acoustic conditions or to meet special preferences of the user regarding the listening habits.

The enhancement is achieved by mixing the ambient sound (AS) with supplementary audio content, called the assistive live listening signal (ALLS), and presenting this individual live mix through in-ear headphones. The AS is picked up by microphones placed at the outside of the headphones close to the listener’s ear canal entrance and the ALLS, e.g. a dry live mix, can be received through a low-latency wireless link (e.g. future mobile communication technology) from a central mixing unit. Since the headset microphones capture the traditional acoustical PA loudspeaker playback, a temporal offset between microphone signals and supplementary content can cause irritating echoes or comb filtering effects. Thus, this time delay has to be estimated and compensated for. In this paper we consider the supplementary audio content to be the same stereo mix that is fed to the PA loudspeaker system. Furthermore, we focus on an indoor scenario. Although the presented method can be transferred to other augmented reality audio applications, the described audience service was the particular motivation of this work.

The problem of time delay estimation (TDE) occurs in numerous applications of acoustic signal processing. It is an important tool in microphone array processing e.g. for acoustic source localization [3, 4]. Several methods for estimating the temporal offset between two or more audio signals have been proposed in the past. Some of the most popular approaches are adaptive eigenvalue decomposition, the maximum-likelihood method or cross-correlation-based methods [5]. For the latter, both time domain and frequency domain implementations exist.

One crucial issue in TDE is the presence of reverberation [6]. While some approaches deal with this problem inherently, others use a pre-filtering of the input signals to obtain a more accurate and robust estimate in reverberant environments. A cepstral pre-filtering [7] as well as a SNR-dependent frequency weighting [8, 9] have been proposed.

From the methods mentioned above we chose the most computationally efficient ones for our application for being able to easily implement them on a real-time DSP platform later on. Since the algorithm in our application shall detect and compensate for time delays of up to 80 ms at a sampling rate of 48 kHz, the number of correlation lags L was chosen to be 4096. This enormous number of lags means a significant advantage for frequency domain methods, which reduce the computational complexity to $O(L \cdot log(L))$. Thus, two frequency domain-based techniques, the regular frequency domain cross-correlation (FD-CC) and the generalized cross-correlation phase transform (GCC-PHAT), were examined with respect to suitability in this application. In particular, a highpass pre-filtering of the input signals for both algorithms was evaluated. These results represent a first step and will be followed by further studies.

Overview of Algorithms

Frequency Domain Cross-Correlation

The fundamental idea behind the cross-correlation based TDE techniques is to find a time lag p that maximizes the the cross-correlation function $r_{x1,x2}(p)$ of the two correlated signals x_1 and x_2. The cross-correlation function is defined as follows:

$$r_{x1,x2}(p) = E\left[x_1(n)x_2(n+p)\right]. \tag{1}$$

In actual implementations we are usually working with a time-averaged estimate of the cross-correlation $\hat{r}_{x1,x2}(p)$. This estimate can be computed in both time and frequency domain. Due to the computational complexity reasons mentioned above, we decided to use the frequency domain implementation.

The cross-correlation can be estimated as

$$\hat{r}_{x1,x2}(p) = IFFT\{X_1(k)X_2^*(k)\} \tag{2}$$

where

$$X_i(k) = FFT\{x_i(n)\}, \quad i = 1, 2 \tag{3}$$

is the FFT of a set of observation samples $x_i(n)$ of the corresponding signal x_i.

The time delay $\hat{\tau}_{1,2}$ can be obtained as

$$\hat{\tau}_{1,2} = \arg\max_p \ \hat{r}_{x1,x2}(p). \tag{4}$$

Please note that $\hat{r}_{x1,x2}(p)$ is a biased cross-correlation estimate.

The downside of using the cross-correlation for TDE applications is its susceptibility to errors caused by reverberation and tonal components in the input signals.

GCC-PHAT

In contrast to the FD-CC, the so called generalized cross-correlation phase transform (GCC-PHAT) additionally uses a weighting function $\Phi(k)$ in the frequency domain which makes the GCC-PHAT insensitive to reverberation and tonality in music and speech signals. This property made the GCC-PHAT well known and widely used in TDE applications, but it needs more computational resources than the FD-CC.

The weighting function is given as

$$\Phi(k) = \frac{1}{|G_{x1,x2}(k)|} \tag{5}$$

where

$$G_{x1,x2}(k) = E\left[X_1(k)X_2^*(k)\right] \tag{6}$$

is the cross spectrum of x_1 and x_2. In practical systems it has to be replaced by its instantaneous values

$$\hat{G}_{x1,x2}(k) = X_1(k)X_2^*(k). \tag{7}$$

We can calculate the desired estimate as

$$\hat{r}_{x1,x2}^{PHAT}(p) = IFFT\left\{\frac{X_1(k)X_2^*(k)}{|X_1(k)X_2^*(k)|}\right\} \tag{8}$$

and obtain the time delay by

$$\hat{\tau}_{1,2} = \arg\max_p \ \hat{r}_{x1,x2}^{PHAT}(p). \tag{9}$$

Pre-filtering of signals

Music and speech signals usually contain strong tonal components especially in lower and middle frequency bands. These periodic components cause several secondary peaks in the signals' auto- and cross-correlation functions on both sides of the primary peak. This effect makes it more difficult to find the primary peak. Furthermore, a reverberation in the room additionally causes secondary peaks since it adds numerous delayed copies of the original signal. The same as for the tonal components applies here: Reverberation is strongest in the lower and middle frequency bands because sound energy at higher frequencies is absorbed by surfaces to a higher degree than at low frequencies.

We make use of this a-priori knowledge about the signals and the room by applying a highpass filtering to both the ALLS and the AS signals. For the actual implementation a 2nd order Butterworth highpass filter was used. The choice of the cutoff frequency will be discussed in the experiments.

Of course, a filtering of signals not only alters the magnitude but also the phase response which can be a problem when relying on the phase response for further calculations. Since we apply the same filtering to both signals their relative phase response remains unchanged.

As it will be shown later in the experiments, the high-pass pre-filtering of the input signal results in an improvement for both the FD-CC and the GCC-PHAT algorithm.

Results of Evaluation

Model Setup for Simulations

After performing some simulation tests with synthesized signals from speech and music samples with binaural room impulse responses (BRIRs) from a database, we conducted binaural audio recordings in a real room and used those signals for our actual experiments. The room chosen for the recordings was a rather large lecture hall (approx. 220 seats, with gallery level, $T_{60,mean} = 1.53\,\text{s}$) which was considered sufficiently reverberant. The ALLSs were played back through a stereo loudspeaker setup in the front of the lecture room. The AS signals were recorded with a binaural microphone prototype (blocked ear canal, Sennheiser KE3 microphone capsules) worn by a dummy head (Neumann KU 100) located in the rear third of the room in a half left position (see Fig. 1). Later ALLS and AS signals were used for offline simulations of the chosen TDE methods.

Four different audio tracks were used as ALLS signals: two speech samples (one male, one female) and two music samples. Of course, these samples cannot be considered representative for all signals the service might be used with, but they can be considered typical signals in possible application scenarios of the audience service. The following signals were used:

- **Sample 1:** Male Speech
  Man reading a text from a book,

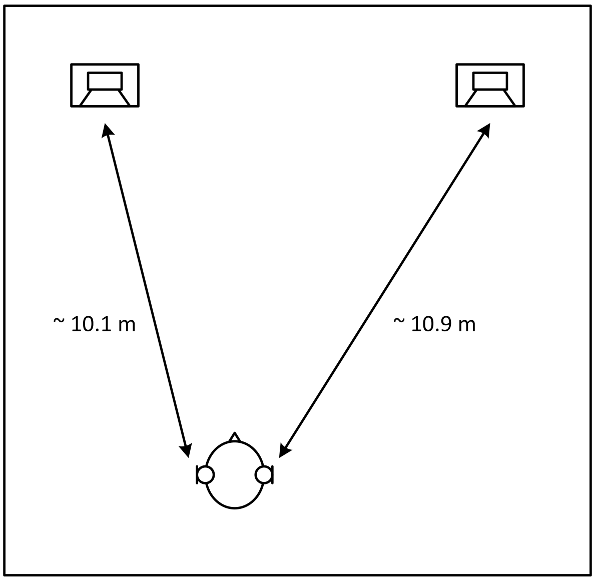

Figure 1: Recording scenario (schematic illustration).

- **Sample 2:** Female Speech
  Woman reading a news text,
- **Sample 3:** Pop/Rock Music
  Mark Knopfler "Boom, Like That",
- **Sample 4:** Pop/Rock Music
  Porcupine Tree "Trains".

To be able to analyze the estimation performance of the algorithms with and without stereo channel crosstalk the signals were recorded with both stereo channels active and with only one channel active.

Before the signals were used for offline simulations they were normalized with regard to the RMS value of the respective waveform.

Cutoff Frequency for Pre-Filtering

For the evaluation of the cutoff frequency offline simulations were conducted. One minute long samples of ALLS and AS of the four audio tracks were pre-filtered with a highpass filter of different cutoff frequencies. The signals were then fed to the blockwise estimation by the FD-CC and GCC-PHAT. For each block the obtained solution was compared to the original time delay whereas a deviation of ± 3 samples was tolerated. For this evaluation the time delay from left speaker to left ear microphone without stereo crosstalk had to be estimated. The curves in Figs. 2 to 5 show the ratio of correct estimates (RCE) over the applied cutoff frequency of the pre-filtering.

For all audio samples a cutoff frequency of 8 kHz shows better performance results than all its preceding values, which is considered an optimum case here. The block-by-block TDE of the two algorithms is followed by a recursive averaging which further improves the estimation results.

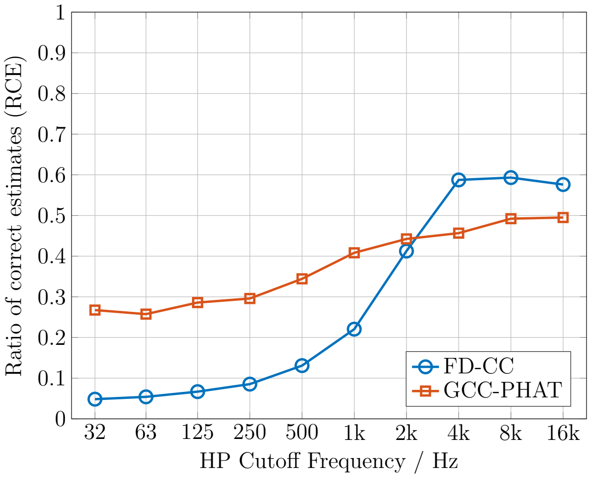

Figure 2: Sample 1: RCE over cutoff frequency.

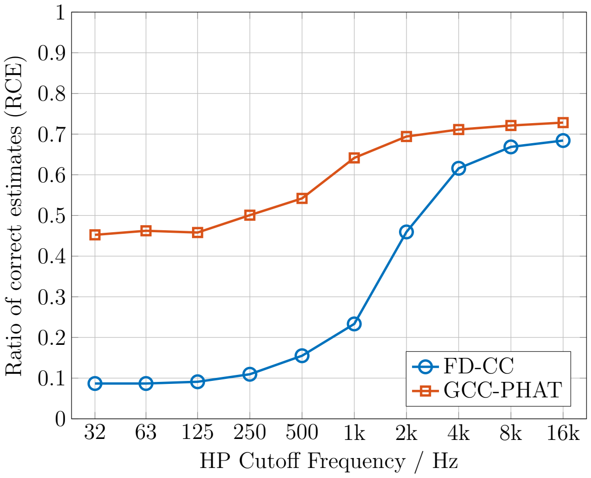

Figure 3: Sample 2: RCE over cutoff frequency.

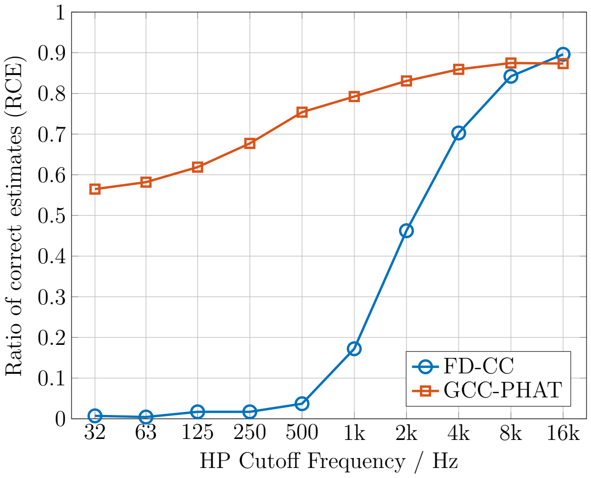

Figure 4: Sample 3: RCE over cutoff frequency.

Summary

Our experiments have shown that in our application of the Assistive Live Listening Service a highpass pre-filtering of the input signals eliminates the most disturbing effects like tonality and reverberation for the FD-CC.

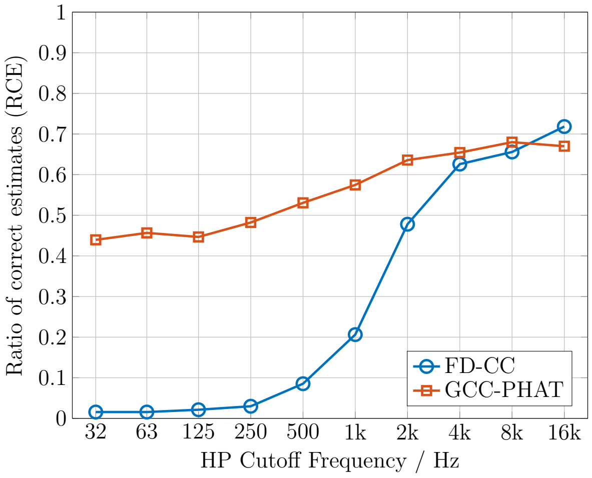

Figure 5: Sample 4: RCE over cutoff frequency

This allows to achieve equivalent estimation results compared to the pre-filtered GCC-PHAT, while being less computational intensive. Based on our tests with four audio samples and one room the optimum cutoff frequency was chosen to be 8 kHz. A limiting aspect is the fact that our methods rely on sufficient signal energy above 8 kHz. So it might not be applicable to all kinds of music or in any arbitrary room.

In future research work the presented algorithms have to be examined more thoroughly. A well known issue for most TDE algorithms is a decreasing performance in the presence of noise or uncorrelated signal energy in the input signals. Furthermore the estimation performance for the stereo crosstalk scenario has to be evaluated.

Acknowledgements

The authors would like to thank their partners in the PMSE-xG project [1] and also the German Federal Ministry of Transport and Digital Infrastructure (BMVI) for the funding of the project (reference number: DG12/8331.3-4 PMSE-xG).

References

[1] PMSE-xG Project Webpage, `http://pmse-xg.de/`, 2017.

[2] Hupke, R., Nophut, M., Preihs, S., and Peissig, J., “5G-Enabled Augmented Audience Services for Live Events,” in *Fortschritte der Akustik : DAGA 2018, München: 9.-22. März 2018 : 44. Jahrestagung für Akustik*, 2018.

[3] Jingdong Chen, Yiteng Huang, and Benesty, J., “Time Delay Estimation via Multichannel Cross-Correlation,” in *Proceedings. (ICASSP ’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, IEEE, 2005.

[4] Thyssen, J., Pandey, A., and Borgstrom, B. J., “A novel Time-Delay-of-Arrival estimation technique for multi-microphone audio processing,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015.

[5] Huang, Y., Benesty, J., and Chen, J., *Acoustic MIMO Signal Processing*, Signals and Communication Technology, Springer Berlin Heidelberg, 2006.

[6] Champagne, B., Bedard, S., and Stephenne, A., “Performance of time-delay estimation in the presence of room reverberation,” *IEEE Transactions on Speech and Audio Processing*, 1996.

[7] Stephenne, A. and Champagne, B., “Cepstral prefiltering for time delay estimation in reverberant environments,” in *1995 International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1995.

[8] Kang, H.-g., Graczyk, M., and Skoglund, J., “On pre-filtering strategies for the GCC-PHAT algorithm,” in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2016.

[9] Grondin, F. and Michaud, F., “Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[10] Knapp, C. and Carter, G., “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976.